K近邻模型

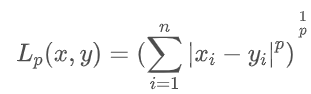
（1）K近邻

k近邻法使用的模型实际上对应于对特征空间的划分。其基本思想可以用这样一句俗语来解释——“近朱者赤，近墨者黑”。

具体来说，在使用K近邻算法的时候，提前定义好样本之间的距离与K值的前提下，对于任意一个新的样本，将其分类为与该样本距离最近的K个样本中类别最多的那个类别。有点裁判给运动员进行判分的那种感觉。

1. 距离度量

很多计算都是基于距离来进行计算和评判的，度量两个样本之间是否相似等属性，一般都是使用距离来度量的。距离度量方法如下：



当P = 1的时候，这种距离称为曼哈度距离。

当P =2的时候，这种距离称为欧氏距离

当P =∞的时候，这种距离称为各个坐标距离的最大值距离。

1. K值选择

k 值的选择会对K近邻法的结果产生重大影响．

较小的 k 值，就相当于用较小的邻域中的训练实例进行预测，缺点是“学习”的估计误差会增大，容易发生过拟合．

如果选择较大的 k值，就相当于用较大邻域中的训练实例进行预测． k 值的增大就意味着整体的模型变得简单，会学习到太多无用的信息。

所以在实际的项目中我们先取一个比较小的K值，然后再使用交叉验证的方法来进行选择。